

Distributed Data Mining (DDM) for NASA Databases and Streams

Hillol Kargupta

Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County

Krishnamoorthy Sivakumar

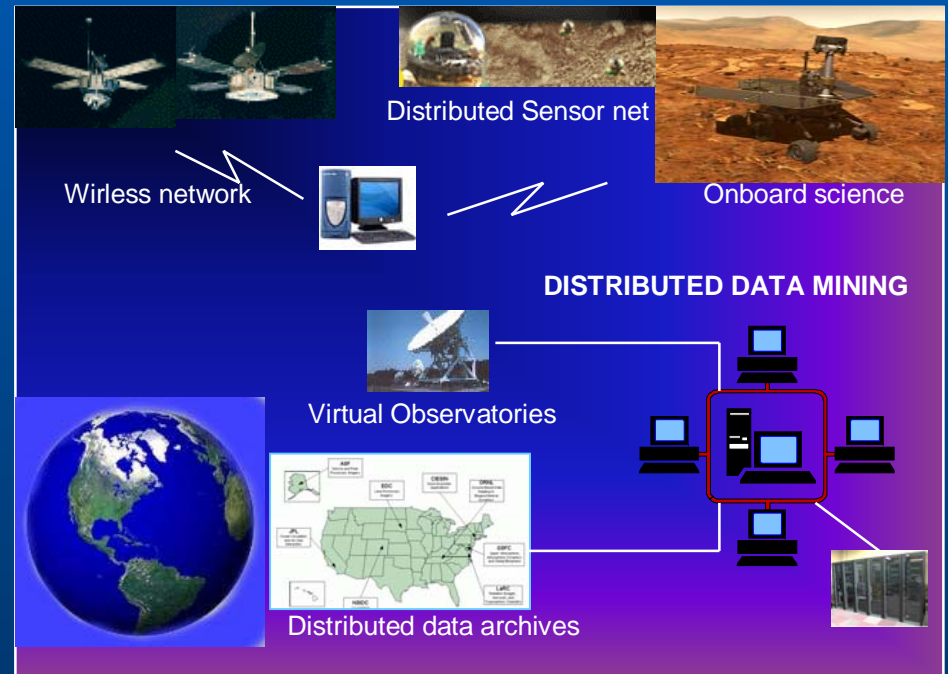
School of Electrical Engineering and Computer Science
Washington State University

Outline

- Project Overview
- Technical Approach
 - Distributed Mining of decision trees
 - Distributed Mining of Bayesian networks
- Experimentation with data from
 - Virtual Observatories data
 - NASA/NOAA data

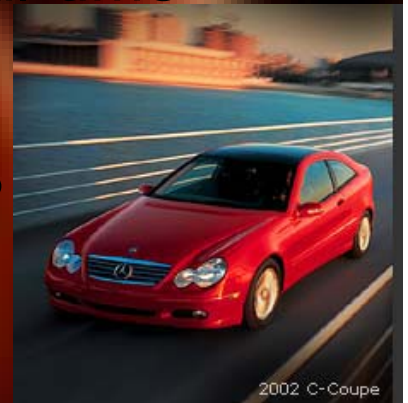
Project Overview

- Goal: Analyzing distributed heterogeneous data by properly utilizing distributed resources.
- Contributions:
 - Distributed decision tree and Bayesian net learning algorithms.
 - Algorithms for monitoring distributed data streams.
 - Mining NASA/NOAA AVHRR data and the virtual observatory data.



Broader Impacts

- Mining Databases from distributed sites
 - Counter-terrorism, bioinformatics
- Monitoring Multiple time critical data streams
 - Monitoring vehicle data streams in real-time
 - Onboard science
- Analyzing Lightweight sensor webs
 - Limited network bandwidth
 - Limited power supply
- Preserving privacy
 - Security/Safety related applications



Why Bother?

x1	x2
4	1
4	5
6	8
1	4
7	1

x1	x3
4	5
1	6
5	9
2	10
7	4

x1	x2	x3
4	1	5
4	5	5
1	4	6
7	1	4
...

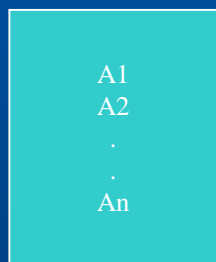
- (Left) Data table at site 1.
- (Middle) Data table at site 2.
- (Right) Joined data table (based on the shared feature x_1) needed for centralized data mining systems.
- Problems:
 - Construction of the join is computationally expensive
 - Supporting repeated queries (e.g. for streams) may be too expensive for the communication-bandwidth.

Example: Network of Virtual Observatories

- The Sloan Digital Sky Survey (SDSS) and the 2MASS All-Sky Survey.
- Five filters from SDSS and three filters from 2MASS.
- Compute the color ratios.
- Features are measured on a logarithmic scale.
- Compute pairwise differences of object features, where the features across surveys

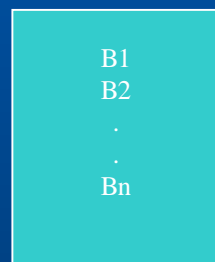
Distributed Inner Product Computation

Site 1



$Z_{1,k}$

Site2



$Z_{2,k}$



■ Site 1 computes Z_{1k}

- $Z_{1k} = A_1.J_1 + \dots + A_n.J_n$

- $J_i \in \{+1, -1\}$

■ Site 2 calculates Z_{2k}

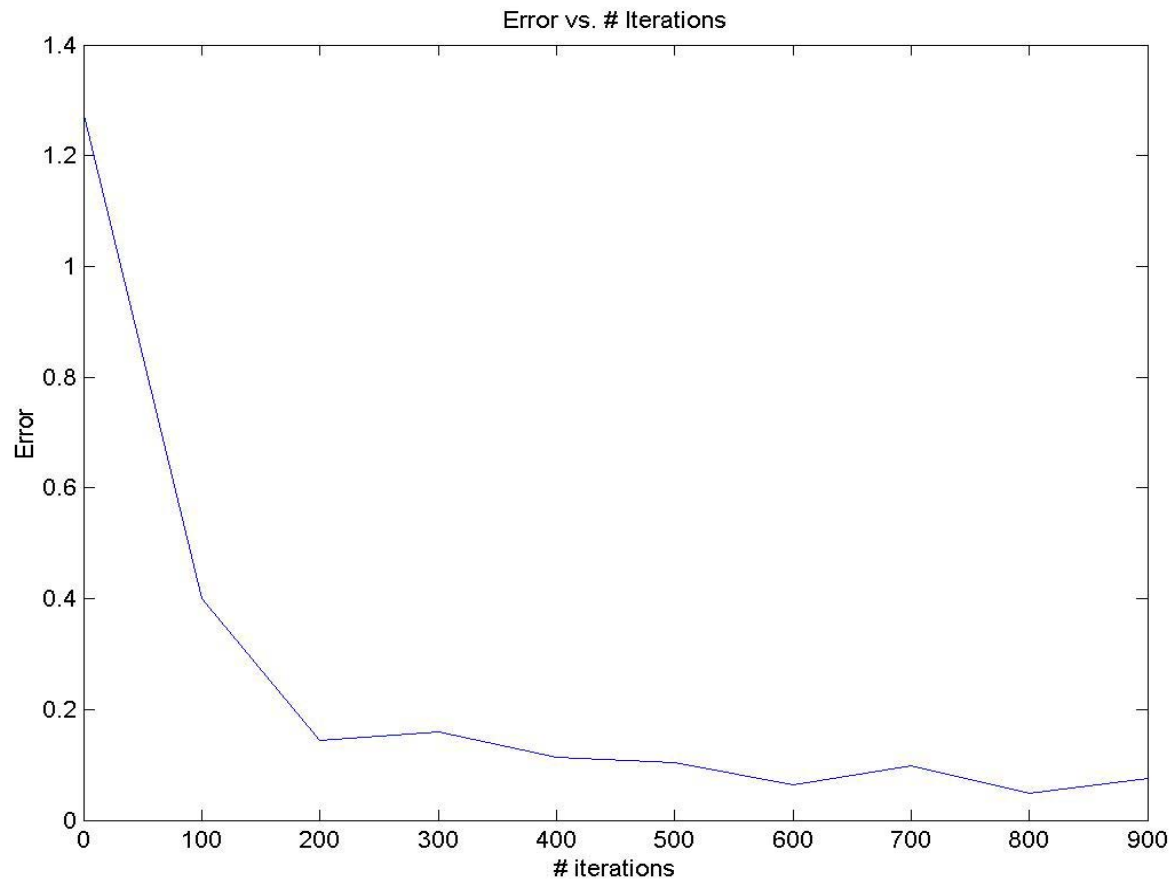
- $Z_{2k} = B_1.J_1 + \dots + B_n.J_n$

■ Compute $z_{1k} \cdot z_{2k}$ for a few times and take the average

Heterogeneous DDM and Decision Trees

- Distributed Randomized Inner Product (DRIP) computation
- Computing information gain using DRIP.
- Information gain computation can be posed as an inner product computation problem.

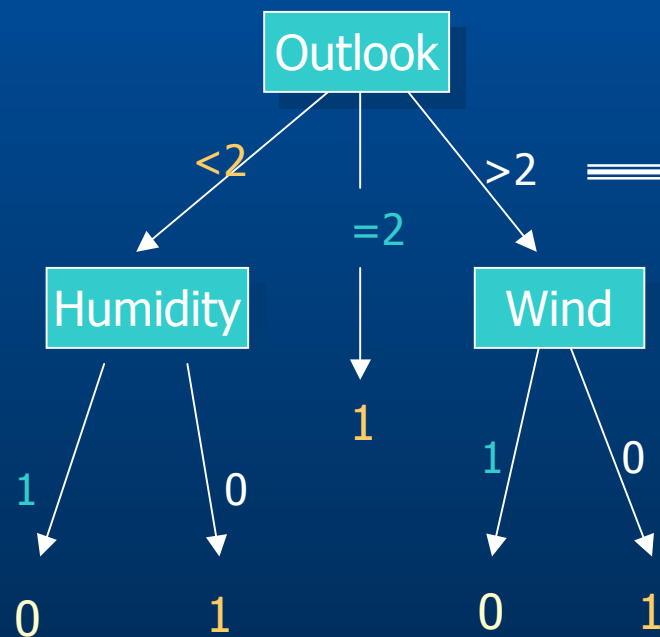
Relative Error vs. Communication Cost



Ensemble of Decision Trees & Homogeneous DDM

- Ensemble-based approaches are popular for handling homogeneous DDM applications:
 - Bagging
 - Arcing
 - SEA (Streaming Ensemble Algorithm)
- Problems:
 - Large ensembles are difficult to interpret
 - Expensive cost of communication

Fourier spectrum of a Decision tree



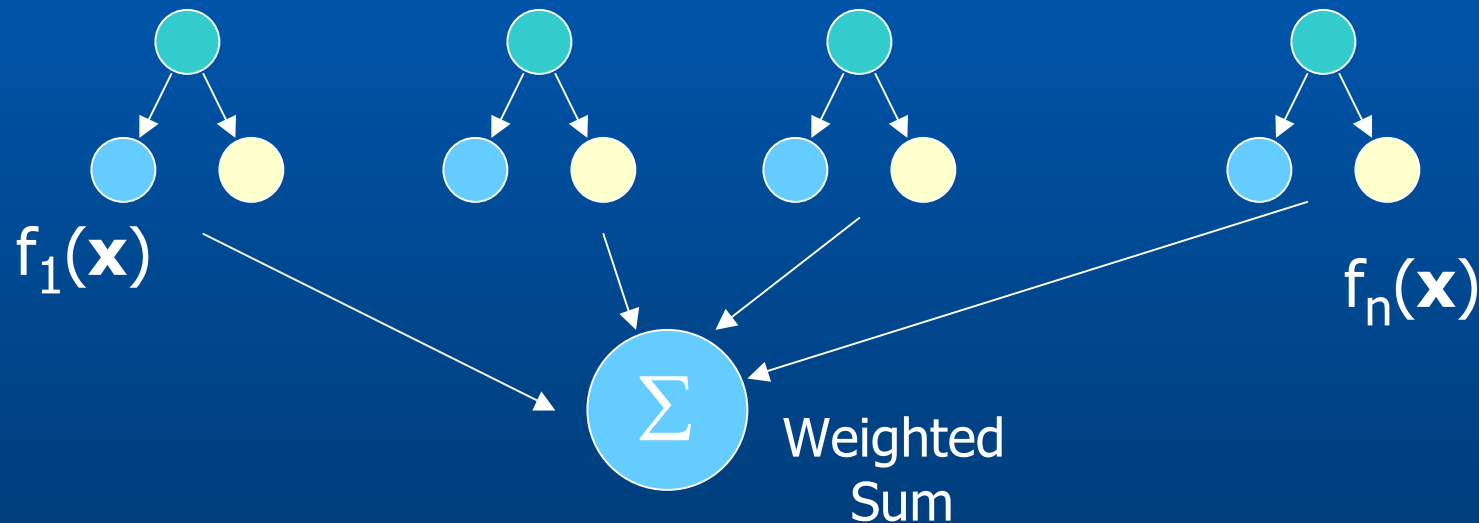
Fourier Coefficient (FC)

$$f(x) = \sum_j w_j \psi_j(x)$$

partition

Fourier Basis Function

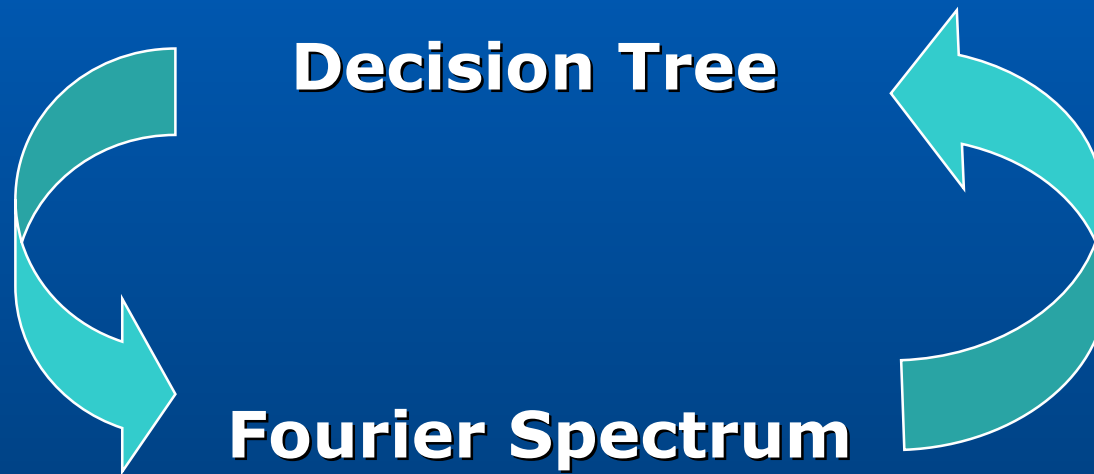
Fourier Spectrum of an Ensemble Classifier



- $$F(\mathbf{x}) = a_1 f_1(\mathbf{x}) + a_2 f_2(\mathbf{x}) + \dots + a_n f_n(\mathbf{x}).$$
$$= A_1 \sum_{j_1} \mathbf{w}_{j_1} \Psi_{j_1}(\mathbf{x}) + \dots + a_n \sum_{j_n} \mathbf{w}_{j_n} \Psi_{j_n}(\mathbf{x}).$$
$$= \sum_{\mathbf{j}} \mathbf{w}_{\mathbf{j}} \Psi_{\mathbf{j}}(\mathbf{x})$$

\mathbf{j} is union of all j_i .

Fourier Spectrum and Decision Trees



- Developed efficient algorithms to
 - Compute Fourier spectrum of decision tree
 - Compute tree from the Fourier spectrum
- Orthogonal Decision trees
 - Redundancy free
 - Stability analysis

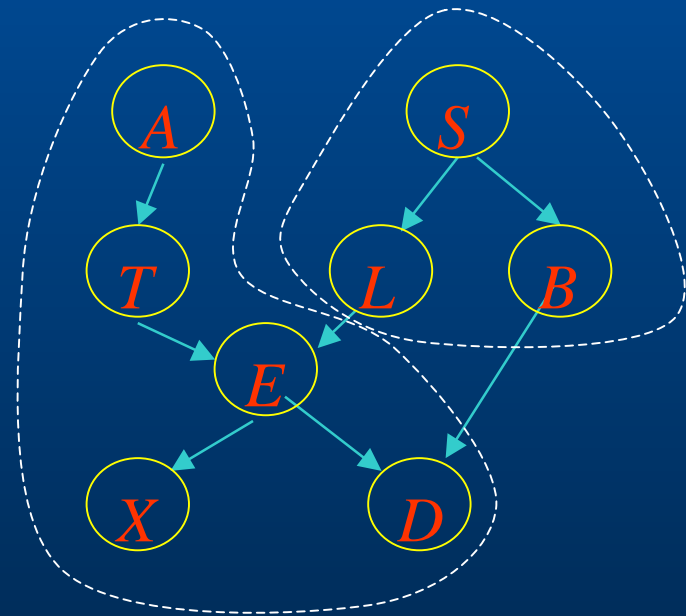
Visualization of Decision Trees



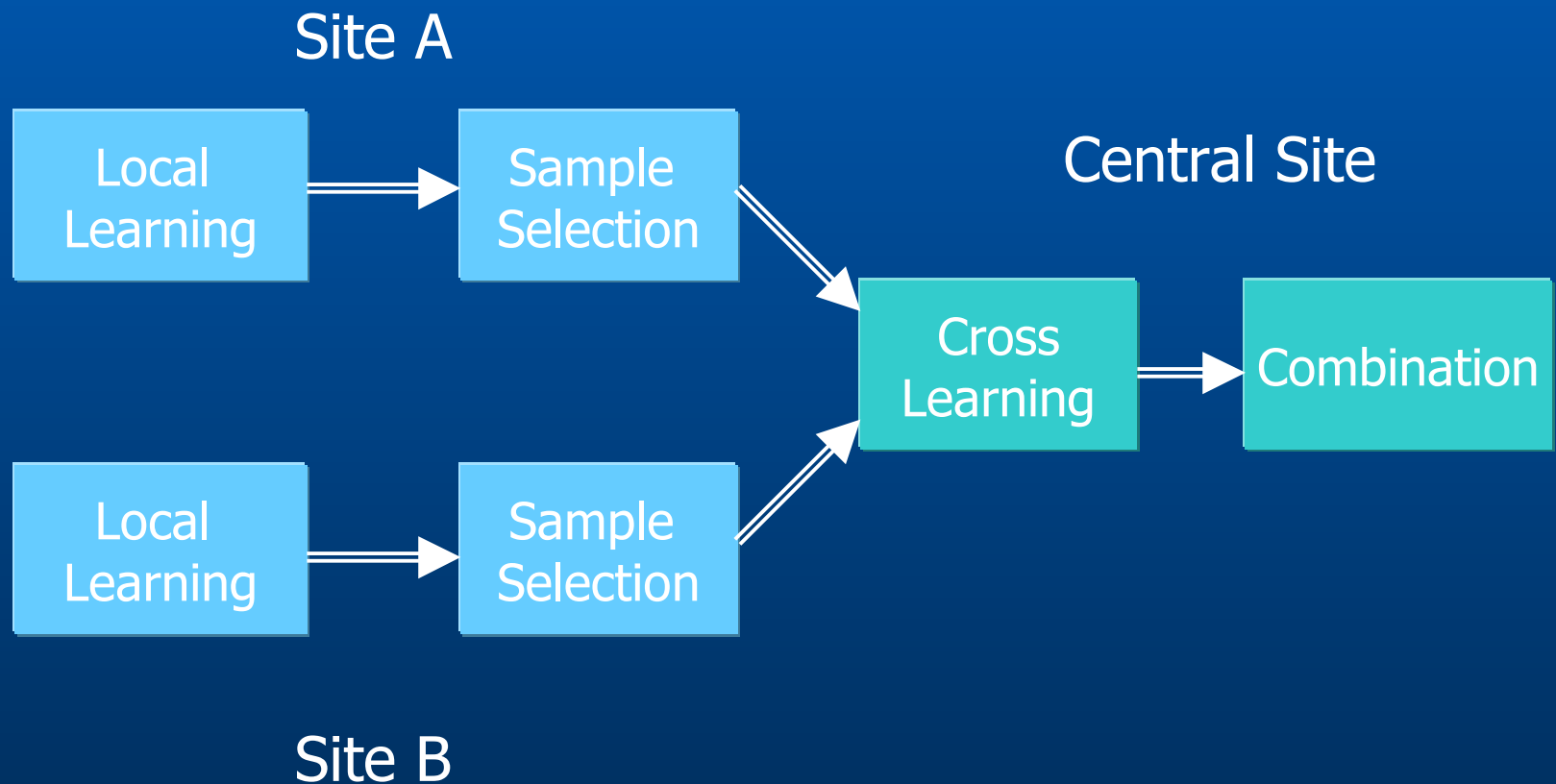
- FC are color-coded in accordance to the magnitude.
- Brighter spots are more significant coefficients.
- On clicking, partition corresponding to the coefficient is displayed.

Distributed BN Learning

- A Bayesian network (BN) is a probabilistic graph model.
- Two problems: Structure and Parameter learning.

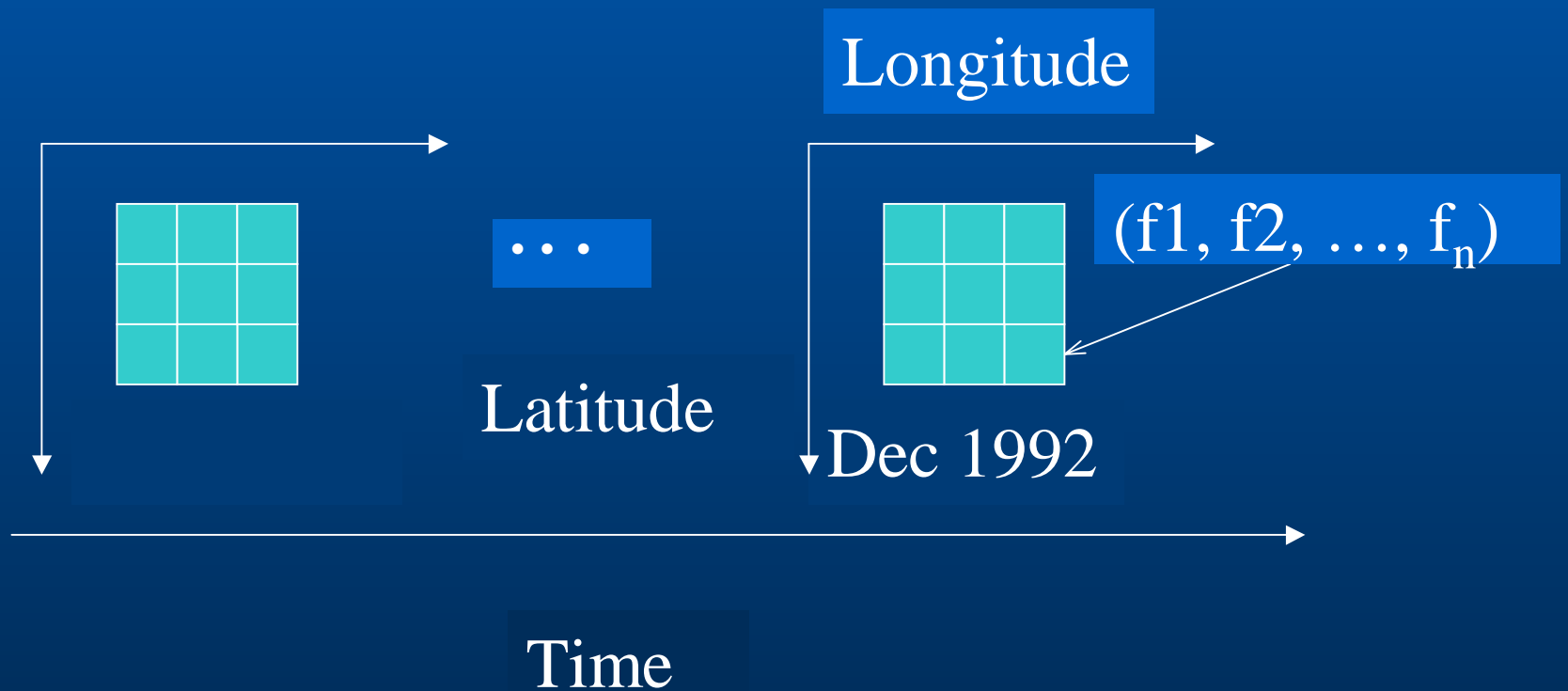


Collective BN Learning



NASA DAO/NOAA AVHRR Pathfinder Data Model

- Multi-dimensional time series data



Preprocessing

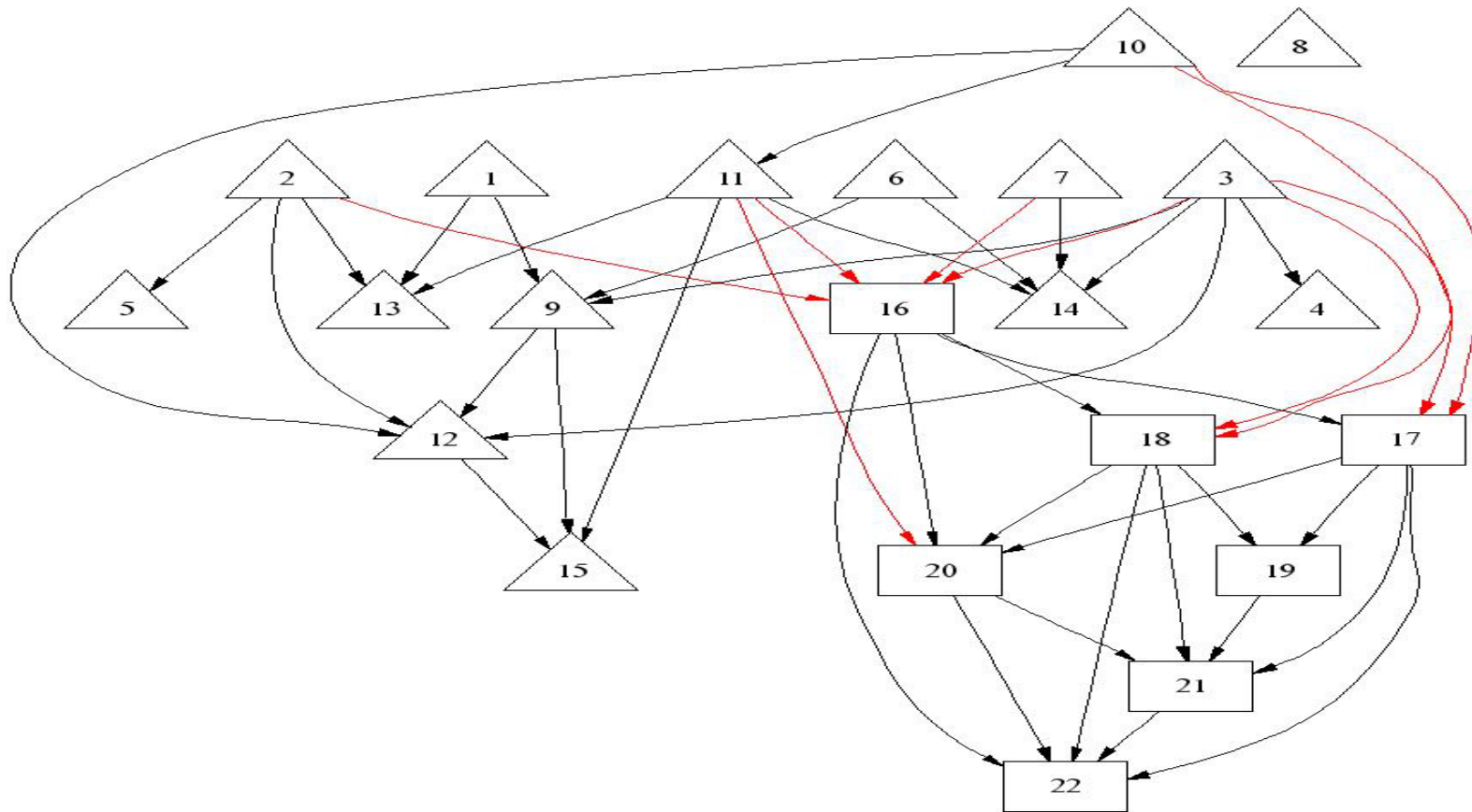
- Feature Selection
- Data Coordination
- Clustering: Segment grid points into local homogenous regions.
- Z score normalization
- Quantization

Bayesian network Learning Results

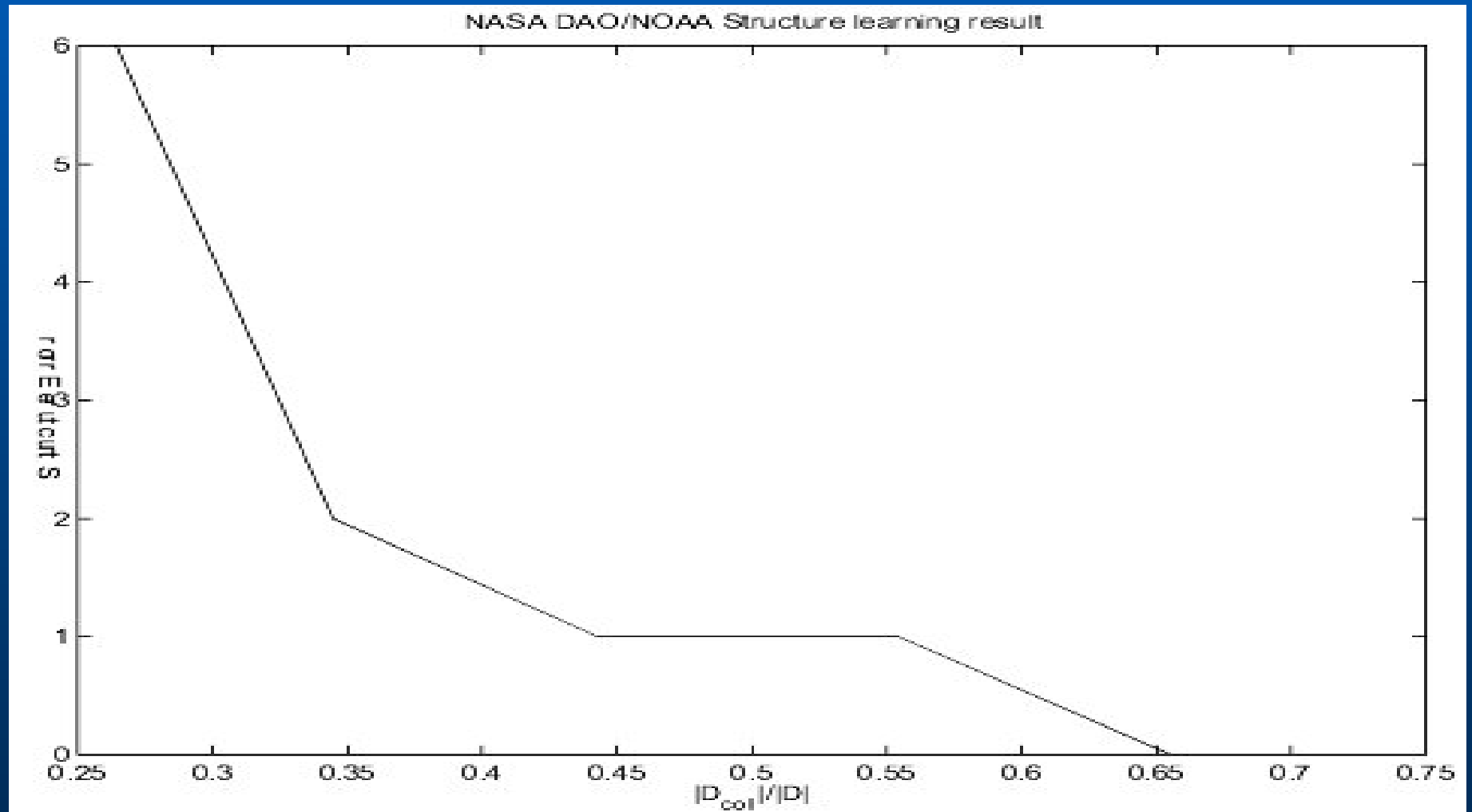
- Compare the Bayesian Networks:
 - B_{cntr} learnt using centralized method.
 - B_{coll} learnt using collective method.
- Metric: structure error = Number of missing links + Number of extra links.

Result

- B_{cntr} – 64 local links and 9 cross links.



Collective Learning Result



Collective Learning

- With 35% samples, get 7 correct cross links and 1 extra links.
- With 45% samples, get 8 correct cross links and 1 extra links.
- With 66% samples, No error.

Web site

- <http://www.cs.umbc.edu/~hillol/nasap.html>

Selected Publications

- Kargupta, H. and Park, B. (2004). A Fourier Spectrum-Based Approach to Represent Decision Trees for Mining Data Streams in Mobile Environments. IEEE Transaction on Knowledge and Data Engineering, Volume 16, Number 2, pages 216--229.
- Chen, R., Sivakumar, K., and Kargupta, H. Collective Mining of Bayesian Networks from Distributed Heterogeneous Data. (accepted) Knowledge and Information Systems, 2002.
- Chen, R. and Sivakumar, K. A New Algorithm for Learning Parameters of a Bayesian Network from Distributed Data. (To appear) Proceedings of the IEEE International Conference on Data Mining, 2002, IEEE Press.
- Chen, R., Sivakumar, K., and Kargupta, H. Distributed Web Mining using Bayesian Networks from Multiple Data Streams. Proceedings of the IEEE International Conference on Data Mining, 75--82. IEEE Press.
- Chen, R., Sivakumar, K., and Kargupta, H. An Approach to Online Bayesian Learning from Multiple Data Streams, Proceedings of the Workshop on Mobile and Distributed Data Mining, PKDD2001.
- Kargupta, H. and Park, B. Mining Decision Trees from Data Streams in a Mobile Environment. Proceedings of the IEEE International Conference on Data Mining, 281--288. IEEE Press.
- Park, B. and Kargupta, H. Constructing Simpler Decision Trees from Ensemble Models Using Fourier Analysis, Proceedings of ACM SigMod DMKD'02 Workshops, Madison, WI (To appear).
- Ayyagari, R. and Kargupta, H. A Resampling Technique for Learning the Fourier Spectrum of Skewed Data, Proceedings of ACM SigMod DMKD'02 Workshops, Madison, WI (To appear).

Feature Selection

- We used as many features as possible.
- Features with following characteristics were dropped.
 - Many missing values
 - Multi-layer features
 - Almost deterministic features
- Used 15 DAO and 7 NOAA features

Feature List

- 1 – Cldfrfc, 2 – Evaps, 3 – Olr, 4 – Osr, 5 – Pbl, 6 – preacc, 7 – qint, 8 – radlwg, 9 – radswg, 10 – t2m, 11 – tg, 12 – ustar, 13 – vintuq, 14 – vintvq, 15 – winds, 16 – asfts, 17 – olrcs_day, 18 – olrcs_night, 19 – olrts_day, 20 – olrts_night, 21 – tcf_day, 22 – tcf_night
- Features 1 - 15 is from NASA DAO and 16 - 22 is from NOAA

Coordination and Clustering

- **Coordination:** re-grid the NOAA dataset into DAO format.
- **Spatio-temporal Clustering:** Segment datasets into local homogenous regions in spatial and temporal domain.
- Each cluster is modeled using a Bayesian network.

Spatio-temporal Clustering

- **Temporal clustering:** choose same month data.
- **Spatial clustering**
 - Average the data from same month. Get one frame of data in spatial domain.
 - Clustering: k-mean, fuzzy c-mean, and EM.

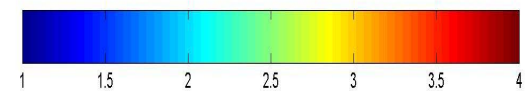
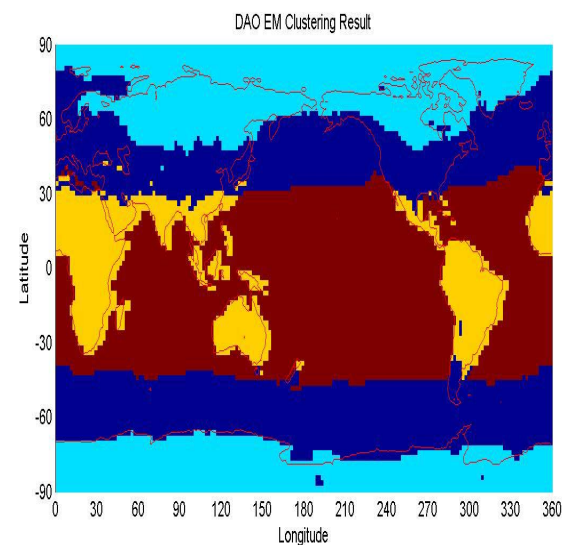
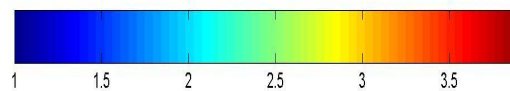
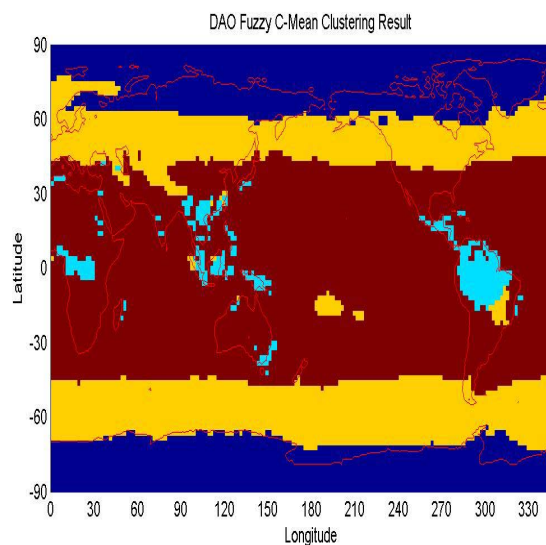
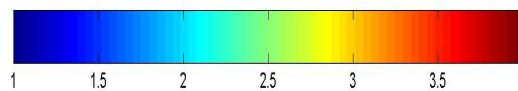
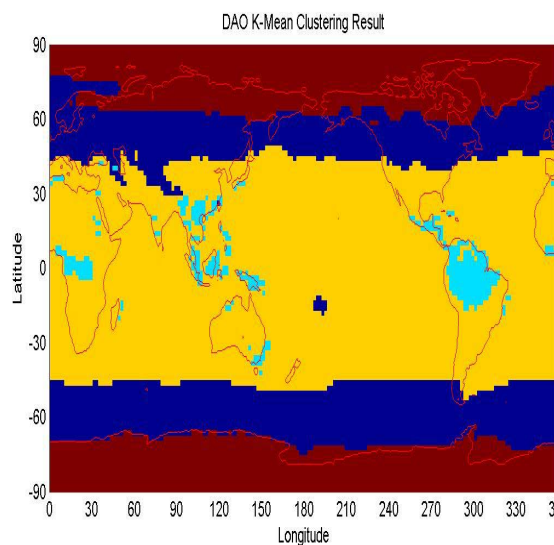
NASA DAO

- Subset of the DAO monthly mean data set: 26 features
- Temporal Coverage: March 1980 - November 1993
- Temporal Resolution: All gridded values are monthly means
- Spatial Coverage: Global
- Spatial Resolution: 2 degree x 2 degree, grid point data (180 x 91 values per level)

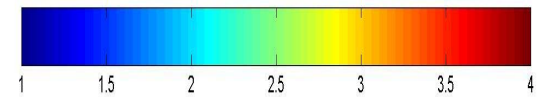
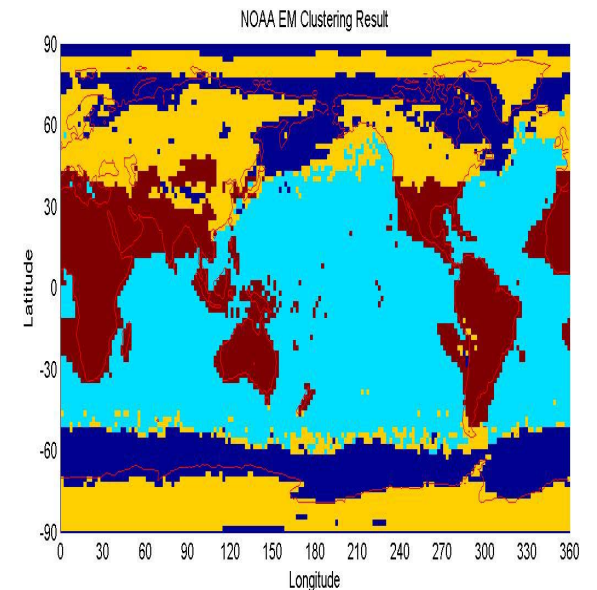
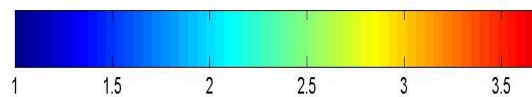
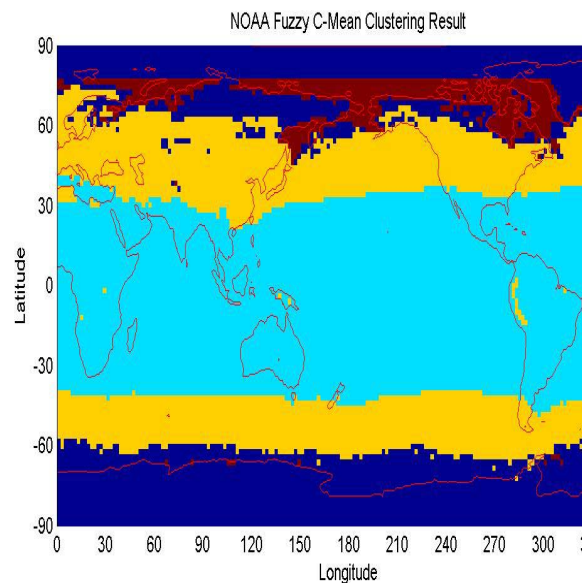
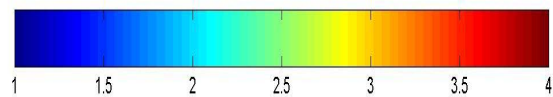
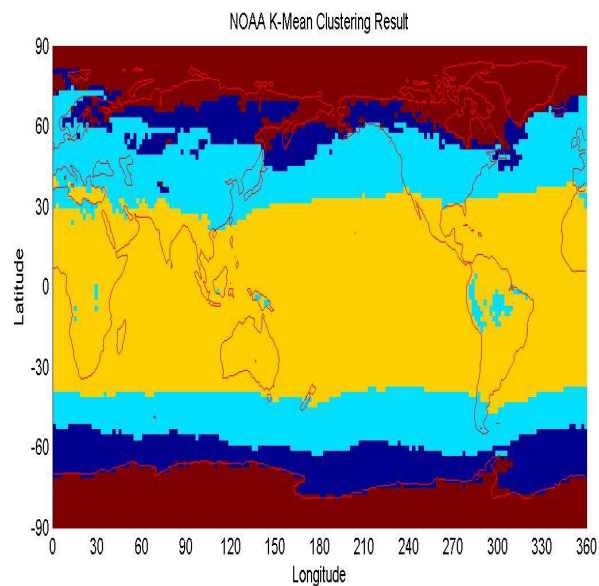
NOAA AVHRR Pathfinder

- Product of NOAA AVHRR Pathfinder: 9 features
- Temporal Coverage: July 1981 - November 2000
- Temporal Resolution: All gridded values are monthly means
- Spatial Coverage: Global
- Spatial Resolution: 1 degree x 1 degree, grid point data (360 x 180 values per level, proceeding west to east and then north to south)

Clustering Results: DAO



Clustering Results: NOAA



Preprocessing

- **Clustering:** Chose a cluster that roughly corresponds to the rectangular region from (170W, 60S) to (90W, 0)
- **Z score normalization** $d_z = \frac{d - \mu}{\sigma}$
- **Quantization:** Discretize the continuous feature value into discrete levels based on its histogram.
- After above steps, we get 12 datasets, one for each month (aggregated over years 1983-1992).

Quantization Results

